Variance Regularized Counterfactual Risk Minimization via Variational Divergence Minimization

Hang Wu¹, May D. Wang¹
¹Wallace H. Countler Department of Biomedical Engineering, Georgia Institute of Technology and Emory University, Atlanta, GA, 30332, USA.

Abstract

Off-policy learning, the task of evaluating and improving policies using historic data collected from a logging policy, is important because on-policy evaluation is usually expensive and has adverse impacts. One of the major challenges of off-policy learning is to derive counterfactual estimators that also have low variance and thus low generalization error. In this work, inspired by learning bounds for importance sampling problems, we present a new counterfactual learning principle for off-policy learning with bandit feedbacks. Our method regularizes the generalization error by minimizing the distribution divergence between the logging policy and the new policy, and removes the need for iterating through all training samples to compute sample variance regularization in prior work. With neural network policies, our end-to-end training algorithms using variational divergence minimization showed significant improvement over conventional baseline algorithms and is also consistent with our theoretical results.

Introduction

Off-policy learning refers to evaluating and improving a deterministic policy using historic data collected from a stationary policy, which is important because in real-world scenarios on-policy evaluation is oftentimes expensive and has adverse impacts. For instance, evaluating a new treatment option, a clinical policy, by administering it to patients requires rigorous human clinical trials, in which patients are exposed to risks of serious side effects. As another example, an online advertising A/B testing can incur high cost for advertisers and bring them few gains. Therefore, we need to utilize historic data to perform off-policy evaluation and learning that can enable safe exploration of the hypothesis space of policies before deploying them.

There has been extensive studies on off-policy learning in the context of reinforcement learning and contextual bandits, including various methods such as Q learning (Sutton and Barto 1998), doubly robust estimator (Dudík et al. 2014), self-normalized (Swaminathan and Joachims 2015b), etc. A recently emerging direction of off-policy learning involves the use of logged interaction data with bandit feedback. However, in this setting, we can only observe limited feedback, often in the form of a scalar reward or loss, for every action; a larger amount of information about other possibilities is never revealed, such as what reward
we could have obtained had we taken another action, the best action we should have take, and the relationship between the change in policy and the change in reward. For example, after an item is suggested to a user by an online recommendation system, although we can observe the user’s subsequent interactions with this particular item, we cannot anticipate the user’s reaction to other items that could have been the better options.

Using historic data to perform off-policy learning in bandit feedback case faces a common challenge in counterfactual inference: *How do we handle the distribution mismatch between the logging policy and a new policy and the induced generalization error?* To answer this question, (Swaminathan and Joachims 2015a) derived the new counterfactual risk minimization framework, that added the sample variance as a regularization term into conventional empirical risk minimization objective. However, the parametrization of policies in their work as linear stochastic models has limited representation power, and the computation of sample variance regularization requires iterating through all training samples. Although a first-order approximation technique was proposed in the paper, deriving accurate and efficient end-to-end training algorithms under this framework still remains a challenging task.

Our contribution in this paper is three-fold:

1. By drawing a connection to the generalization error bound of importance sampling (Cortes, Mansour, and Mohri 2010), we propose a new learning principle for off-policy learning with bandit feedback. We explicitly regularize the generalization error of the new policy by minimizing the distribution divergence between it and the logging policy. The proposed learning objective automatically trade off between empirical risk and sample variance.

2. To enable end-to-end training, we propose to parametrize the policy as a neural network, and solves the divergence minimization problem using recent work on variational divergence minimization (Nowozin, Cseke, and Tomioka 2016) and Gumbel soft-max (Jang, Gu, and Poole 2016) sampling.

3. Our experiment evaluation on benchmark datasets shows significant improvement in performance over conventional baselines, and case studies also corroborates the soundness of our theoretical proofs.

**Background**

**Problem Framework**

We first review the framework of off-policy learning with logged bandit feedback introduced in (Swaminathan and Joachims 2015a). A policy maps an input $x \in \mathcal{X}$ to a structured (discrete) output $y \in \mathcal{Y}$. For example, the input $x$ can be profiles of users, and we recommend movies of relevance to the users as the output $y$; or in the reinforcement learning setting, the input is the trajectory of the agent, and the output is the action the agent should take in the next time point. We use a family of stochastic policies, where each policy defines a posterior distribution over the output space given the input $x$, parametrized by some $\theta$, i.e., $h_{\theta}(y|x)$. Note that here a distribution which has all its probability density mass on one
action corresponds to a deterministic policy. With the distribution \( h(y | x) \), we take actions by sampling from it, and each action \( y \) has a probability of \( h(y | x) \) being selected. In the discussion later, we will use \( h \) and \( h(y | x) \) interchangeably when there will not create any confusion.

In online systems, we observe feedbacks \( \delta(x, y; y^*) \) for the action \( y \) sampled from \( h(Y | x) \) by comparing it to some underlying ‘best’ \( y^* \) that was not revealed to the system. For example, in recommendation system, we can use a scalar loss function \( \delta(x, y; y^*) \) \( \rightarrow [0, L] \), with smaller values indicating higher satisfaction with recommended items.

The risk of a policy \( h(Y | x) \) is defined as

\[
R(h) = \mathbb{E}_{x \sim \mathbb{P}(X), y \sim h(Y | x)}[\delta(x, y)]
\]

, and the goal of off-policy learning is to find a policy with minimum risk on test data.

In the off-line logged learning setting, we only have data collected from a logging policy \( h_0(Y | x) \), and we aim to find an improved policy \( h(Y | x) \) that has lower risks \( R(h) < R(h_0) \). Specifically, the data we will use will be

\[
\mathcal{D} = \{ x_i, y_i, \delta_i = \delta_i(x_i, y_i), p_i = h_0(y_i | x_i) \}, i = 1, ..., N,
\]

where \( \delta_i \) and \( p_i \) are the observed loss feedback and the logging probability (also called propensity score), and \( N \) is the number of training samples.

Two main challenges are associated with this task: 1) If the distribution of a logging policy is skewed towards a specific region of the whole space, and doesn’t have support everywhere, feedbacks of certain actions cannot be obtained and improvement for these actions is not possible as a result. 2) since we cannot compute the expectation exactly, we need to resort to empirical estimation using finite samples, which creates generalization error and needs additional regularization.

A vanilla approach to solve the problem is propensity scoring approach using importance sampling (Rosenbaum and Rubin 1983), by accounting for the distribution mismatch between \( h \) and \( h_0 \). Specifically, we can rewrite the risk w.r.t \( h \) as the risk w.r.t \( h_0 \) using an importance reweighting:

\[
R(h) = \mathbb{E}_{x \sim \mathbb{P}(X), y \sim h(Y | x)}[\delta(x, y)] = \mathbb{E}_{x \sim \mathbb{P}(X), y \sim h_0(Y | x)} \left[ \frac{h(y | x)}{h_0(y | x)} \delta(x, y) \right]
\]

With the collected historic dataset \( \mathcal{D} \), we can estimate the empirical risk \( \hat{R}_{\mathcal{D}}(h) \), short as \( \hat{R}(h) \)

\[
\hat{R}(h) = \frac{1}{N} \sum_{i=1}^{N} h(y_i | x_i) \frac{h(y_i | x_i)}{h_0(y_i | x_i)} \delta_i(x_i, y_i)
\]
Counterfactual Risk Minimization

(Swaminathan and Joachims 2015a) pointed out several flaws with the vanilla approach, namely, not being invariant to loss scaling, large and potentially unbounded variance. To regularize the variance, the authors proposed a regularization term for sample variance derived from empirical Bernstein bounds.

The modified objective function to minimize is now:

$$\hat{R}(h) = \frac{1}{N} \sum_{i=1}^{N} u_i + \lambda \sqrt{\frac{Var(\tilde{u})}{N}},$$

where $u_i = \frac{h(y_i|y_i)(y_i|y_i)}{h_0(y_i|y_i)(y_i|y_i)} \delta_i$, $\tilde{u} = \frac{1}{N} \sum_{i=1}^{N} u_i$ is the average of $\{u_i\}$ obtained from training data, and $Var(\tilde{u})$ is the sample variance of $\{u_i\}$.

As the variance term is dependent on the whole dataset, stochastic training is difficult, the authors approximated the regularization term via first-order Taylor expansion and obtained a stochastic optimization algorithm. Despite its simplicity, such first-order approximation neglects the non-linear terms from second-order and above, and introduces approximation errors while trying to reduce the sample variance.

Variance Regularization Objective

Theoretical Motivation

Instead of estimating variance empirically from the samples, which prohibits direct stochastic training, the fact that we have a parametrized version of the policy $h(y|x)$ motivates us to think: can we derive a variance bound directly from the parametrized distribution?

We first note that the empirical risk term $\hat{R}(h)$ is the average loss reweighted by importance sampling function $\frac{h(y|x)}{h_0(y|x)}$, and a general learning bound exist for the second moment of importance weighted loss:

Let $X$ be a random variable distributed according to distribution $\mathcal{P}$ with density $p(x)$, $Y$ be a random variable, and $\delta(x, y)$ be a loss function over $(x, y)$ that is bounded in $[0, L]$. For two sampling distributions of $y$, $h(y|x)$ and $h_0(y|x)$, define their conditional divergence as $d_2(h(y|x) || h_0(y|x); \mathcal{P}(x))$, we have

$$\mathbb{E}_X \sim \mathcal{P}(X), Y \sim h_0(y|x)[u^2(y|x)\delta^2(x, y)]$$

$$\leq L^2 d_2(h(y|x) || h_0(y|x); \mathcal{P}(x))$$

[thm:bound]
The bound is similar to that of (Cortes, Mansour, and Mohri 2010) for a single random variable except that we are working with a joint distribution over $x, y$ here. Detailed proofs can be found in Appendix [thm:bound].

Let $R_h$ be the risk of the new policy on loss function $\delta$, and $\hat{R}_h$ be the empirical risk. We additionally assume the divergence is bounded by $M$, i.e., $d_2(h\|h_0) \leq d_\infty(h\|h_0) = M$

Then with probability at least $1 - \eta$,

$$R(h) \leq \hat{R}(h) + \frac{2LM\log 1/\eta}{3N} + L\sqrt{\frac{2d_2(h\|h_0; x) \log 1/\eta}{N}}$$

[th:bern]

The proof of this theorem is an application of Bernstein inequality and the second moment bound, and detailed proof is in Appendix. This result highlights the bias-variance trade-offs as seen in empirical risk minimization (ERM) problems, where $\hat{R}_h$ approximates the empirical risk/ bias, and the third term characterize the variance of the solution with distribution divergence. It thus motivates us that in bandit learning setting, instead of directly optimizing the reweighted loss and suffer huge variance in test setting, we can try to minimize the variance regularized objectives as

$$\min_{h = h(Y|X)} \hat{R}(h) + \lambda \sqrt{\frac{1}{N} \sum_{i=1}^{m} d_2(h(\cdot|X_i) \delta_i)}$$

$\lambda = \sqrt{2L^2\log 1/\eta}$ is a model hyper-parameter controlling the trade-off between empirical risk and model variance, but we are still faced with the challenge of setting $\lambda$ empirically and the difficulty in optimizing the objective (See Appendix for a comparison). Thus, in light of the recent success of distributionally robust learning, we explore an alternative formulation of the above regularized ERM in the next subsection.

**Robustly Regularized Formulation**

Instead of solving a ‘loss + regularizer’ objective function, we here study a closely related constrained optimization formulation, whose intuition comes from the method of Langaragian mutliplier for constrained optimization.

The new formulation is:

$$\min_h \frac{1}{N} \sum_{i=1}^{m} \frac{h(y_i|x_i)}{p_i} \delta_y$$

s.t. $d_2(h\|h_0; X) \leq \rho/N$, where $\rho$ is a pre-determined constant as the regularization hyper-parameter.
We analyze the generalization property of the minimizer \( h^* \) to the constrained problem: first showing the bound of empirical risk of \( h^* \) compared to that of \( h_0 \), then move to the expectation setting.

The minimizing value of \( \hat{R}(h^*) \) satisfies that

\[
-\left( \frac{2\rho}{N} \sqrt{\text{Var}} \right) \leq \hat{R}(h^*) - \hat{R}(h_0) \leq -\left( \frac{2\rho}{N} \sqrt{\text{Var}} - \frac{2\rho L}{3N} \right) +
\]

[\text{prop:emp_bound}]

where \( \text{Var} = \text{Var}_{h_0} \delta_i(h_i) \) is the empirical variance for the reweighted loss. The results come from algebraic computation using KKT conditions for a Lagrange multiplier argument, and we show the calculation in Appendix.

With the above proposition, we can prove the following bound for a minimizer \( h^* \) to the optimization problem

For a minimizer \( h^* \), we have that with probability at least \( 1 - \exp(-t) \),

\[
R(h^*) \leq R(h_0) - \frac{2(\sqrt{\rho} - \sqrt{t})}{\sqrt{N}} \sqrt{\text{Var}} + \frac{(4 + 6\rho)L}{3N} t + \frac{6L^2}{3N} + \sqrt{\frac{2\text{Var}_{h_0} \delta_i}{N}}
\]

where \( \text{Var} \) is the policy-reweighted loss, and the constant \( \text{Var}_{h_0} \delta_i \) is the variance of the original losses.

We show the detailed proof in Appendix, which combines Bennett’s inequality and the proposition above. These results suggest that the risk of our minimizer is a good surrogate to the best risk plus a variance term, with their difference controlled by the regularization hyper-parameter \( \rho \) and approaches 0 when \( N \to 0 \). Moreover, we can recognize, in some cases, e.g. \( \text{Var} \) is considerably large while \( \rho > t \), we are guaranteed to have an improvement on the losses.

At first glance, the new objective function removes the needs to compute the sample variance in existing bounds, but when we have a parametrized distribution of \( h(y|x) \), and finite samples \( \{x_i, y_i\}_{i=1}^N \) from \( h_0(y|x) \), estimating the divergence function is not an easy task. In the next subsection, we will present how recent f-gan networks for variational divergence minimization (Nowozin, Cseke, and Tomioka 2016) and Gumbel soft-max sampling (Jang, Gu, and Poole 2016) can help solve the task.

**Discussion: Possibility of Counterfactual Learning:** One interesting aspect of our bounds also stresses the need for the stochasticity of the logging policy (Langford, Strehl, and Wortman 2008). For a deterministic logging policy, if the corresponding probability
distribution can only have some peaked masses, and zeros elsewhere in its domain, our intuition suggests that learning will be difficult, as those regions are never explored. Our theory well reflects this intuition in the calculation of the divergence term, the integral of form \( \int y \frac{h(y|x)}{h_0(y|x)} \frac{h(y|x)}{h_0(y|x)} dy \). A deterministic policy has a non-zero measure region of \( h_0(y|x) \) with probability density of \( h(y|x) = 0 \), while the corresponding \( h(y|x) \) can have finite values in the region. The resulting integral results is thus unbounded, and in turn induces an unbounded generalization bound, making counterfactual learning in this case not possible.

**Adversarial Training Algorithm**

**Adversarial Learning of the Divergence**

The derived variance regularized objective requires us to minimize the square root of the conditional divergence, \( d_2(h \| h_0; \mathcal{P}(X)) = \mathbb{E}_x \left[ \frac{h^2}{h_0} \right] dy \).

For simplicity, we can examine the term inside the expectation operation first. With simple calculation, we have

\[
\int y \frac{h(y|x)^2}{h_0(y|x)} dy = \int y h_0(y|x) \left[ \frac{h(y|x)}{h_0(y|x)} \right]^2 - 1 + 1
\]

\[
= D_f(h(y|x) \| h_0(y|x)) + 1,
\]

where \( f(t) = t^2 - 1 \) is a convex function in the domain \( \{ t : t \geq 0 \} \) with \( f(1) = 0 \). Combining with the expectation operator gives a minimization objective of \( D_f(h \| h_0; \mathcal{P}(X)) \) (+1 omitted as constant).

The above calculation draws connection between our divergence and the \( f \)-divergence measure (Nguyen, Wainwright, and Jordan 2010). Follow the \( f \)-GAN for variational divergence minimization method proposed in (Nowozin, Cseke, and Tomioka 2016), we can reach a lower bound of the above objective as

\[
D_f(h(y|x) \| h_0(y|x); \mathcal{P}(X)) \]

\[
= \mathbb{E}_x \left[ \int y \frac{h(y|x)}{h_0(y|x)} \right] dh_0(y|x)
\]

\[
= \mathbb{E}_x \left[ \sup_T \left\{ E_{y \sim h(T(y))} - E_{y \sim h_0}(f^*(T(y))) \right\} \right]
\]

\[
= \sup_T \left\{ \mathbb{E}_x E_{y \sim h(T(x,y))} - \mathbb{E}_x E_{y \sim h_0}(f^*(T(x,y))) \right\}
\]

\[
\geq \sup_T \left\{ \mathbb{E}_x E_{y \sim h(T(x,y))} - \mathbb{E}_x E_{y \sim h_0}(f^*(T(x,y))) \right\}
\]

\[
= \sup_T \left\{ \mathbb{E}_{x,y \sim h(T(x,y))} - \mathbb{E}_{x,y \sim h_0}(f^*(T(x,y))) \right\}
\]

\[
\triangleq F(T, h)
\]

For the second equality, as \( f \) is a convex function and applying Fenchel convex duality \( (f^* = \sup_u (u^T v - f(u))) \) gives the dual formulation. Because the expectation is taken w.r.t to \( x \).
while the supreme is taken w.r.t. all functions $T$, we can safely swap the two operators. We note that the bound is tight when $T_0(x) = f'(h/h_0)$, where $f'$ is the first order derivative of $f$ as $f'(t) = 2t$ (Nguyen, Wainwright, and Jordan 2010).

The third inequality follows because we restrict $T$ to a family of functions instead of all functions. Luckily, the universal approximation theorem of neural networks (Hornik, Stinchcombe, and White 1989) states that neural networks with arbitrary number of hidden units can approximate continuous functions on a compact set with any desired precision. Thus, by choosing the family of $T$ to be the family of neural networks, the equality condition of the second equality can be satisfied theoretically.

The final objective is a saddle point of a function $T(x, y): X \times Y \rightarrow \mathbb{R}$ that maps input pairs to a scalar value, and the policy we want to learn $h(y|x)$ acts as a sampling distribution.

Although being a lower bound with achievable equality conditions, theoretically, this saddle point trained with mini-batch estimation is a consistent estimator of the true divergence (Proof in Appendix).

We use $D_f = \sup_{T} \int T d\mathbb{P} - \int f^*(T) d\mathbb{Q}$ to denote the true divergence, and

$\hat{D}_f = \sup_{T \in T} \int T(x, y) d\mathbb{P} - \int f^*(T(x, y)) d\mathbb{Q}$

the empirical estimator we use, where $\mathbb{P}$ and $\mathbb{Q}$ are the empirical distribution obtained by sampling from the two distribution respectively.

$\hat{D}_f$ is a consistent estimator of $D_f$.

Again, a generative-adversarial approach (Goodfellow et al. 2014) can be applied. Toward this end, we represent the $T$ function as a discriminator network parametrized as $T_w(x, y)$.

We then parametrize the distribution of our policy $h(y|x)$ as another generator neural network $h_\theta(y|x)$ mapping $x$ to the probability of sampling $y$. For structured output problems with discrete values of $y$, to allow the gradients of samples obtained from sampling backpropagated to all other parameters, we use the Gumbel soft-max sampling (Jang, Gu, and Poole 2016) methods for differential sampling from the distribution $h(y|x)$. We list the complete training procedure Alg. [alg:fgan] for completeness. As shown by (Nowozin, Cseke, and Tomioka 2016), when the saddle point solution exists, under some mild conditions, Alg. [alg:fgan] decreases the objective value with a geometric rate.

$D = \{x_i, y_i\}_{i=1}^N$ sampled from logging policy $h_0$; a predefined threshold $D_0$; an initial generator distribution $h_\theta(y|x)$; an initial discriminator function $T_w(x, y)$; max iteration $I$ An optimized generator $h^*_\theta(y|x)$ distribution that has minimum divergence to $h_0$ initialization Sample a mini-batch ‘real’ samples $(x_i, y_i)$ from $D$ Sample a mini-batch $x$ from $D$, and construct ‘fake’ samples $(x_i, \hat{y}_i)$ by sampling $\hat{y}$ from $h_\theta(y|x)$ with Gumbel soft-max Update $w^{t+1} = w^t + \eta_w \partial F(T_w, h_\theta)$ Update $\theta^{t+1} = \theta^t - \eta_\theta \partial F(T_w, h_\theta)$

For our purpose of minimizing the variance regularization term, we can similarly derive a training algorithm, as the gradient of $t \rightarrow \sqrt{t + 1}$ can also be backpropagated.
Training algorithm

With the above two components, we are now ready to present the full treatment of our end-to-end learning for counterfactual risk minimization from logged data. The following algorithm solve the robust regularized formulation and for completeness, training for the original ERM formulation in Sec. [sub:motivation] (referred to co-training version in the later experiment sections) is included in Appendix.

\[ D = \{ x_i, y_i, p_i, \delta_i \}_{i=1}^N \] sampled from \( h_0 \); regularization hyper-parameter \( \rho \), and maximum iteration of divergence minimization steps \( I \), and max epochs for the whole algorithm \( MAX \)

An optimized generator \( h_\theta^*(y|x) \) that is an approximate minimizer of \( R(w) \) initialiation

Sample a mini-batch of \( m \) samples from \( D \)

Estimate the reweighted loss as

\[ \hat{R}_t = \frac{1}{m} \sum_{i=1}^{m} \frac{h_\theta(y_i|x_i)}{p_i} \delta_i \]

and get the gradient as \( g_1 = \partial_\theta \hat{R}_t \)

Update \( \theta_{t+1} = \theta_t - \eta g_1 \)

Call Algorithm [alg:fgan] to minimize the divergence \( D_2(h||h_0; \mathcal{P}(X)) \) with threshold \( = \rho \), and max iter set to \( I \)

The algorithm works in two separate training steps: 1) update the parameters of the policy \( h \) to minimize the reweighted loss 2) update the parameters of the policy/ generator and the discriminator to regularize the variance thus to improve the generalization performance of the new policy.

Experiments

Experiment Setups

For empirical evaluation of our proposed algorithms, we follow the conversion from supervised learning to bandit feedback method (Agarwal et al. 2014). For a given supervised dataset \( \mathcal{D}^* = \{ (x_i, y_i^*) \}_{i=1}^N \), we first construct a logging policy \( h_0(y|x) \), and then for each sample \( x_i \), we sample a prediction \( y_i \sim h_0(y|x_i) \), and collect the feedback as \( \delta(y_i^*, y_i) \). For the purpose of benchmarks, we also use the conditional random field (CRF) policy trained on 5% of \( \mathcal{D}^* \) as the logging policy \( h_0 \), and use hamming loss, the number of incorrectly misclassified labels between \( y_i \) and \( y_i^* \), as the loss function \( \delta \) ((Swaminathan and Joachims 2015a)). To create bandit feedback datasets \( \mathcal{D} = \{ x_i, y_i, \delta_i, p_i \}_{i=1}^N \), each of the samples \( x_i \) were passed four times to the logging policy \( h_0 \) and sampled actions \( y_i \) were recorded along with the loss value \( \delta_i \) and the propensity score \( p_i = h_0(y_i|x_i) \).

In evaluation, we use two type of evaluation metrics for the probabilistic policy \( h(y|x) \). The first is the expected loss (referred to as ‘EXP’ later) \( R(h) = \frac{1}{N_{test}} \sum_{y \sim h(y|x)} \delta(y^*, y) \), a direct measure of the generalization performance of the learned policy. The second is the average hamming loss of maximum a posteriori probability (MAP) prediction \( y_{MAP} = \arg\max h(y|x) \) derived from the learned policy, as MAP is a faster way to generate predictions without the need for sampling in practice. However, since MAP predictions only depend on the regions with highest probability, and doesn’t take into account the diverse of
predictions, two policies with same MAP performance could have very different
generalization performance. Thus, a model with high MAP performance but low EXP
performance might be over-fitting, as it may be centering most of its probability masses in
the regions where $h_0$ policy obtained good performance.

**Benchmark Comparison**

**Baselines**—Vanilla importance sampling algorithms using inverse propensity score (IPS),
and the counterfactual risk minimization algorithm from (Swaminathan and Joachims
2015a) (POEM) are compared, with both L-BFGS optimization and stochastic optimization
solvers. The hyper-parameters are selected by performance on validation set and more
details of their methods can be found in the original paper (Swaminathan and Joachims
2015a).

Neural network policies without divergence regularization (short as NN-NoReg in later
discussions) is also compared as baselines, to verify the effectiveness of variance
regularization.

**Dataset** We use four multi-label classification dataset collected in the UCI machine learning
repo (Asuncion and Newman 2007), and perform the supervised to bandit conversion. We
report the statistics in Table 2 in the Appendix.

For these datasets, we choose a three-layer feed-forward neural network for our policy
distribution, and a two or three layer feed-forward neural network as the discriminator for
divergence minimization. Detailed configurations can be found in the Appendix 6.

For benchmark comparison, we use the separate training version [alg:sep] as it has faster
convergence and better performance (See the section in Appendix for an empirical
comparison).

The networks are trained with Adam (Kingma and Ba 2014) of learning rate 0.001 and 0.01
respectively for the reweighted loss and the divergence minimization part. We used PyTorch
to implement the pipelines and trained networks with Nvidia K80 GPU cards.

Results by an average of 10 experiment runs are obtained and we report the two evaluation
metrics in Table 1. We report the regularized neural network policies with two Gumbel-
softmax sampling schemes, soft Gumbel soft-max (NN-Soft), and straight-through Gumbel
soft-max (NN-Hard).

As we can see from the result, by introducing a neural network parametrization of the
policies, we are able to improve the test performance by a large margin compared to the
baseline CRF policies, as the representation power of networks are often reported to be
stronger than other models. The introduction of additional variance regularization term
(comparing NN-Hard/Soft to NN-NoReg), we can observe an additional improvement in

---

1 Codes for reproducing the results can be found in the link https://github.com/hang-wu/VRCRM.
both testing loss and MAP prediction loss. We observe no significant difference between
the two Gumbel soft-max sampling schemes.

**Effect of Variance Regularization**

To study the effectiveness of variance regularization quantitatively, we vary the maximum
number of iterations \( I \) in Alg. [alg:sep] we take in each divergence minimization sub loop.
For example, ‘NN-Hard-10’ indicates that we use ST Gumbel soft-max and set the
maximum number of iterations to 10. Here we set the thresholds for divergence slightly
larger so maximum iterations are executed so that results are more comparable. We plot the
expected loss in test sets against the epochs average over 10 runs with error bars using the
dataset yeast.

As we can see from the figure, models with no regularization (gray lines in the figure) have
higher loss, and slower convergence rate. As the number of maximum iterations for
divergence minimization increases, the test loss decreased faster and the final test loss is also
lower. This behavior suggests that by adding the regularization term, our learned policies are
able to generalize better to test sets, and the stronger the regularization we impose by taking
more divergence minimization steps, the better the test performance is. The regularization
also helps the training algorithm to converge faster, as shown by the trend.

**Generalization Performance**

Our theoretical bounds implies that the generalization performance of our algorithm
improves as the number of training samples increases. We vary the number of passes of
training data \( x \) was passed to the logging policy to sample an action \( y \), and vary it in the
range \( 2^{[1,2,\ldots,8]} \) with log scales.

When the number of training samples in the bandit dataset increases, both models with and
without regularization have an increasing test performance in the expected loss and reaches a
relatively stable level in the end. Moreover, regularized policies have a better generalization
performance compared to the model without regularization constantly. This matches our
theoretical intuitions that explicitly regularizing the variance can help improve the
generalization ability, and that stronger regularization induces better generalization
performance. But as indicated by the MAP performance, after the replay of training samples
are more than \( 2^4 \), MAP prediction performance starts to decrease, which suggests the models
may be starting over-fitting already.

**Effect of logging policy vs results**

In this section, we discuss how the effect of logging policies, in terms of stochasticity and
quality, will affect the learning performance and additional visualizations of other metrics
can be found in the Appendix.

As discussed before, the ability of our algorithm to learn an improved policy relies on the
stochasticity of the logging policy. To test how this stochasticity affects our learning, we
modify the parameter of \( h_0 \) by introducing a temperature multiplier \( \alpha \). For CRF logging
policies, the prediction is made by normalizing values of \( \omega^T \phi(x,y) \), where \( \omega \) is the model
parameter and can be modified by $\alpha$ with $w \rightarrow aw$. As $\alpha$ becomes higher, $h_0$ will have a more peaked distribution, and ultimately become a deterministic policy with $\alpha \rightarrow \infty$.

We varied $\alpha$ in the range of $2^{[-1,1, \ldots , 8]}$, and report the average ratio of expected test loss to the logging policy loss of our algorithms (Y-axis in Fig 3a, where smaller values indicate a larger improvement). We can see that NN policies are performing better than logging policy when the stochasticity of $h_0$ is sufficient, while after the temperature parameter increases greater than $2^3$, it’s much harder and even impossible (ratio > 1) to learn improved NN policies. We also note here that the stochasticity doesn’t affect the expected loss values themselves, and the drop in the ratios mainly resulted from the decreased loss of the logging policy $h_0$. In addition, comparing within NN policies, policies with stronger regularization have slight better performance against models with weaker ones, which in some extent shows the robustness of our learning principle.

Finally, we discuss the impact of logging policies to the our learned improved policies. Intuitively, a better policy that has lower hamming loss can produce bandit datasets with more correct predictions, however, it’s also possible that the sampling biases introduced by the logging policy is larger, and such that some predictions might not be available for feedbacks. To study the trade-off between better policy accuracy and the sampling biases, we vary the proportion of training data points used to train the logging policy from 0.05 to 1, and compare the performance of our improved policies obtained by in Fig. 3b. We can see that as the logging policy improves gradually, both NN and NN-Reg policies are outperforming the logging policy, indicating that they are able to address the sampling biases. The increasing ratios of test expected loss to $h_0$ performance, as a proxy for relative policy improvement, also matches our intuition that $h_0$ with better quality is harder to beat.

**Conclusion**

In this paper, we showed how divergence regularization can help improve off-policy learning both theoretically via importance sampling bounds and empirically by our adversarial learning algorithms on real-world datasets.

Limitations of the work mainly lies in the need for the propensity scores (the probability an action is taken by the logging policy), which may not always be available. Learning to estimate propensity scores and plug the estimation into our training framework will increase the applicability of our algorithms. For example, as suggested by (Cortes, Mansour, and Mohri 2010), directly learning importance weights (the ratio between new policy probability to the logging policy probability) has comparable theoretical guarantees, which might be a good extension for the proposed algorithm.

Although the work focuses on off-policy from logged data, the techniques and theorems may be extended to general supervised learning and reinforcement learning. It will be interesting to study how this training algorithm can work for empirical risk minimization and what generalization bounds it may have as the future direction of research.
Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This work has been supported in part by grants from the National Science Foundation (NSF) R01CA163256, CDC HHSD2002015F62350B, the Children’s Healthcare of Atlanta, and Microsoft Research. We also thank Dr. Adith Swaminathan for answering questions regarding their paper’s experiments procedure.

Reference:


Figure 1:
Stronger regularization can help obtain faster convergence and better test performance.
Figure 2:
Neural network policies both have increasing performance with increasing number of training data, while models with regularization have faster convergence rate and better performance.
Figure 3:
a) The decreasing stochasticity of $h_0$ makes it harder to obtain an improved NN policy, and our regularization can help the model be more robust and achieve better generalization performance.
b) As $h_0$ improves, the models constantly outperform the baselines, however, the difficulty is increasing with the quality of $h_0$. Note: more visualizations of other metrics can be found in the appendix 6.
Figure 4:
Results from different training schemes suggest that separately minimizing reweighted loss and divergence is better compared to training the two losses together.
Figure 5:
As the logging policy becomes more deterministic, NN policies are still able to find improvement over h0 in a) expected loss and b) loss with MAP predictions. c) We cannot observe a clear trend in terms of the performance of MAP predictions. We hypothesize it results from that h0 policy already has good MAP prediction performance by centering some of the masses. While NN policies can easily pick up the patterns, it will be difficult to beat the baselines. We believe this phenomenon worth further investigation.
Figure 6:

(a) As the quality of the logging policy increases, NN policies are still able to find improvement over $h_0$ in expected loss.

and (b) (c) For MAP predictions, however, it will be really difficult for NN policies to beat if the logging policy was already exposed to full training data and trained in a supervised fashion.
Table 1:

Benchmark Comparison Results.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Scene MAP</th>
<th>Scene EXP</th>
<th>Yeast MAP</th>
<th>Yeast EXP</th>
<th>TMC MAP</th>
<th>TMC EXP</th>
<th>LYRL MAP</th>
<th>LYRL EXP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logging Policy $h_0$ (5% data CRF)</td>
<td>1.069</td>
<td>1.887</td>
<td>3.255</td>
<td>5.485</td>
<td>4.995</td>
<td>5.053</td>
<td>1.047</td>
<td>1.949</td>
</tr>
<tr>
<td>NN-NoReg</td>
<td>1.465</td>
<td>1.981</td>
<td>4.705</td>
<td>1.706</td>
<td>1.724</td>
<td>0.247</td>
<td>0.263</td>
<td></td>
</tr>
<tr>
<td>NN-Hard</td>
<td>1.303</td>
<td>1.463</td>
<td>3.047</td>
<td>3.788</td>
<td>1.694</td>
<td>1.720</td>
<td>0.248</td>
<td><strong>0.255</strong></td>
</tr>
<tr>
<td>NN-Soft</td>
<td>1.347</td>
<td>1.457</td>
<td>3.097</td>
<td>3.789</td>
<td>1.683</td>
<td><strong>1.707</strong></td>
<td>0.247</td>
<td><strong>0.255</strong></td>
</tr>
<tr>
<td>POEM</td>
<td>1.169</td>
<td><strong>1.169</strong></td>
<td>4.238</td>
<td>4.508</td>
<td>4.611</td>
<td>4.505</td>
<td>1.169</td>
<td>1.306</td>
</tr>
<tr>
<td>IPS(Stochastic)</td>
<td>1.291</td>
<td>1.291</td>
<td>4.090</td>
<td>4.605</td>
<td>2.812</td>
<td>2.737</td>
<td>1.149</td>
<td>1.479</td>
</tr>
<tr>
<td>POEM(Stochastic)</td>
<td>1.322</td>
<td>1.323</td>
<td>4.140</td>
<td>4.570</td>
<td>3.601</td>
<td>3.561</td>
<td>1.237</td>
<td>1.237</td>
</tr>
<tr>
<td>Supervised Learning (NN)</td>
<td>0.943</td>
<td>2.238</td>
<td>3.101</td>
<td>4.300</td>
<td>1.530</td>
<td>3.786</td>
<td>0.217</td>
<td>0.519</td>
</tr>
<tr>
<td>Supervised Learning (CRF)</td>
<td>1.110</td>
<td>1.423</td>
<td>2.807</td>
<td>4.047</td>
<td>1.344</td>
<td>1.241</td>
<td>0.240</td>
<td>0.437</td>
</tr>
</tbody>
</table>

*Proc Mach Learn Res. Author manuscript; available in PMC 2020 August 11.*
### Table 2: Dataset Statistics

<table>
<thead>
<tr>
<th>Name</th>
<th># Features</th>
<th># Labels</th>
<th># Train</th>
<th># Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yeast</td>
<td>103</td>
<td>14</td>
<td>1500</td>
<td>917</td>
</tr>
<tr>
<td>Scene</td>
<td>294</td>
<td>6</td>
<td>1211</td>
<td>1196</td>
</tr>
<tr>
<td>TMC</td>
<td>30438</td>
<td>22</td>
<td>21519</td>
<td>7077</td>
</tr>
<tr>
<td>LYRL</td>
<td>47236</td>
<td>4</td>
<td>23149</td>
<td>781265</td>
</tr>
</tbody>
</table>