Boosting Black Box Variational Inference

Francesco Locatello * (flocatello@tuebingen.mpg.de)[ab] · Gideon Dresdner * [a] · Rajiv Khanna [c] · Isabel Valera [b] · Gunnar Rätsch [a]

[a] ETH · [b] MPI · [c] TEXAS The University of Texas at Austin

BIOMEDICAL INFORMATICS : We are hiring!

Overview

By relaxing the constraints on boosting variational inference algorithms, we are able to develop a new black box variational inference algorithm which also enjoys convergence guarantees.

- ✔ Relaxed assumption for convergence
- ✔ Black box boosting subroutine
- ✔ Stopping conditions

Boosting VI

Variational inference

$$\min_{q \in \mathcal{Q}} D^{KL}(q \| p_{\boldsymbol{x}})$$

Limitation

- ✤ Selection of an appropriate Q to trade expressivity with tractability

Boosting Variational Inference

$$\min_{q \in \text{conv}(\mathcal{Q})} D^{KL}(q \| p_x)$$

Properties

- ✔ Convex objective
- ✔ Convex constraint set
- ✔ Bounded Curvature
- ✔ Provable rate: $\varepsilon_t \leq \frac{2\left(\frac{1}{\delta} C_{f,\mathcal{A}} + \varepsilon_0\right)}{\delta t + 2}$

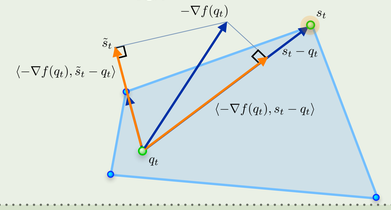
Affine Invariant Frank-Wolfe [3]

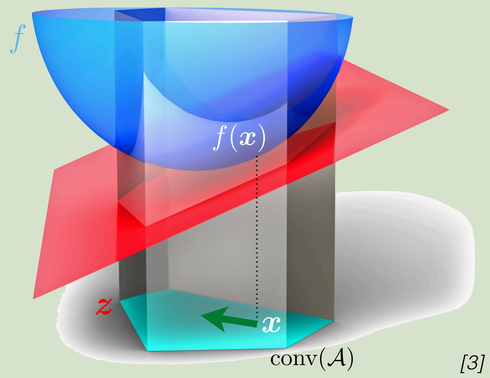
1: **init** $q^0 \in \text{conv}(\mathcal{Q})$, and $\mathcal{S} := \{\mathbf{q}_0\}$
2: **for** $t = 0 \dots T$
3: Find $s^t := (\text{Approx-})\text{LMO}_{\mathcal{Q}}(\nabla D(q^t))$
4: Variant 0: $\gamma = \frac{2}{t+2}$
5: Variant 1: $\gamma = \arg\min_{\gamma \in [0,1]} D((1-\gamma)q^t + \gamma s^t)$
6: $q^{t+1} := (1-\gamma)q^t + \gamma s^t$
7: Variant 2: $\mathcal{S} = \mathcal{S} \cup s^t$
8: $q^{t+1} = \arg\min_{q \in \mathcal{S}} D(q)$
9: **end for**

Boosting Subroutine

Linear Minimization Oracle:

$$\langle \nabla f(q_t), \tilde{s}_t - q_t \rangle \leq \delta \min_{s \in \mathcal{A}} \langle \nabla f(q_t), s - q_t \rangle \qquad \delta \in (0, 1]$$





[3]

Curvature

Theorem:

$$C_{f,\mathcal{A}} := \sup_{\substack{s \in \mathcal{A},\, q \in \text{conv}(\mathcal{A}) \\ \gamma \in [0,1] \\ y = q + \gamma(s-q)}} \frac{2}{\gamma^2} D^{KL}(y \| q)$$

is bounded for the KL divergence if the parameter space of the densities in $\mathcal{A}$ is bounded.

Solving the LMO the naive way:

- ✤ Naive degenerate solution
- ✔ Need to constrain infinity norm

Dealing with the Infinity Norm

Lemma: A density with bounded infinity norm has entropy bounded from below. The converse is true for many of the distributions which are commonly used in VI (for example Gaussian, Cauchy and Laplace).

Optimizing the RELBO

The Entropy Constraint

$$\arg\min_{\substack{s \in \bar{\mathcal{A}} \\ H(s) \geq -M}} \left\langle s, \log\left(\frac{q^t}{p}\right) \right\rangle$$

$$\arg\min_{s \in \bar{\mathcal{A}}} \left\langle s, \log\left(\frac{q^t}{p}\right) \right\rangle + \lambda\left(-H(s) - M\right)$$

Equivalent KL Minimization

$$\arg\min_{s \in \overline{\mathcal{A}}} D^{KL}\left(s \middle\| \frac{1}{Z} \sqrt[\lambda]{\frac{p_x}{q}}\right)$$

RELBO

$$\text{RELBO}(s, \lambda) := \mathbb{E}_s[\log p] - \lambda \mathbb{E}_s[\log s] - \mathbb{E}_s[\log q^t].$$

- ✔ New iterate close to the posterior
- ✔ New iterate far from the old iterate
- ✔ New iterate has high entropy

Duality Gap

$$g(q) := \max_{s \in \text{conv}(\mathcal{A})} \langle q - s, \log \frac{q}{p} \rangle \geq D^{KL}(q \| p) - D^{KL}(q^\star \| p)$$

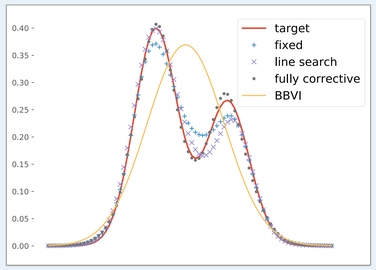
- ✔ An exact oracle gives a bound on the error for free
- ✤ We cannot compute the exact oracle
- ✔ Can use it to monitor the training

Failure Modes

- ✤ Line search is tricky in high dimension
- ✤ Symmetric posterior
- ✤ Fitting the tails
  - ✤ If one component is already good
  - ✔ Duality gap helps

Experiments

Proof of Concept



Bayesian Logistic Regression

$$w \sim \mathcal{N}(0, 1)$$

$$\mathbf{y} \sim \text{Bernoulli}(\text{logit}^{-1}(w^\top X))$$

ChemReact

	Train LL	Test AUROC
Boosting BBVI (Laplace)	-.677 ± 0.002	**0.794 ± 0.005**
BBVI Edward (Laplace)	-0.681 ± 0.003	0.781 ± 0.012
BBVI Edward (Gaussian)	-0.671 ± 0.002	0.790 ± 0.009
Line Search Boosting VI	-2.808	0.6377
Fixed Step Boosting VI	-3.045	0.6193
Norm Corrective Boosting VI	-2.725	0.6440

eICU Collaborative Research

	Train LL	Test AUROC
Boosting BBVI (Laplace)	**-0.195 ± 0.007**	**0.844 ± 0.006**
BBVI Edward (Laplace)	-0.200 ± 0.032	0.838 ± 0.016

Probabilistic Matrix Factorization

$$U, V \sim \mathcal{N}(0, 1); \quad X \sim \mathcal{N}(UV^\top, \sigma)$$

CBCL Faces

	BBVI MSE	Boosting BBVI MSE
D=3	0.0184 ± 0.001	**0.0139 ± 0.44e-04**
D=5	0.0187 ± 0.001	**0.0137 ± 0.53e-04**
D=10	0.0188 ± 0.001	**0.0135 ± 0.52e-04**
	BBVI Test LL	Boosting BBVI Test LL
D=3	-0.9363 ± 0.6e-3	**-0.9354 ± 0.3e-3**
D=5	**-0.9391 ± 0.6e-3**	-0.9393 ± .4e-3
D=10	**-0.9468 ± 0.3e-3**	-0.9492 ± .001

- *http://github.com/ratschlab/boosting-bbvi*

References

[1] *Boosting Variational Inference: an Optimization Perspective.* Locatello, Khanna, Ghosh, Rätsch
[2] *Boosting Variational Inference.* Fangjian Guo, Xiangyu Wang, Kai Fan, Tamara Broderick, David B. Dunson
[3] *Revisiting Frank-Wolfe: Projection-Free Sparse Convex Optimization.* Martin Jaggi